The Costs of Science in the Exascale Era

Prof. Dr. Thomas Ludwig

German Climate Computing Centre
Hamburg, Germany
ludwig@dkrz.de

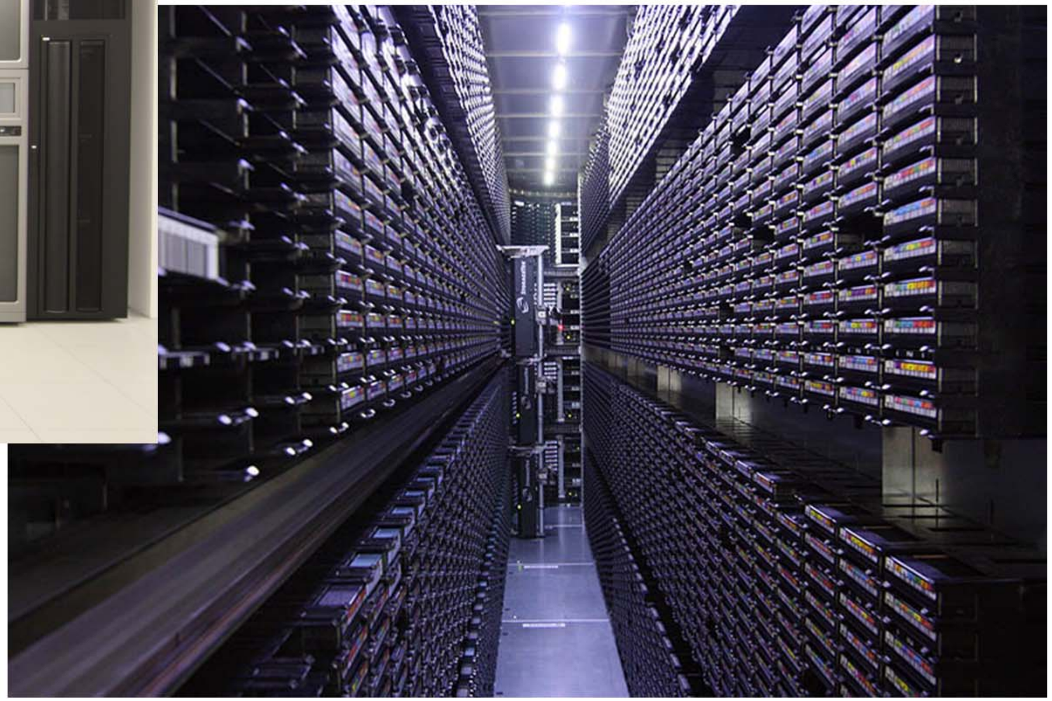
The Terascale and Petascale Era

DKRZ in Hamburg

IBM Power6 Computer System

- Rank 58 in TOP500/Nov10
- 8064 cores, 115 TFLOPS Linpack
- 6PB disks

Sun StorageTek Tape Library



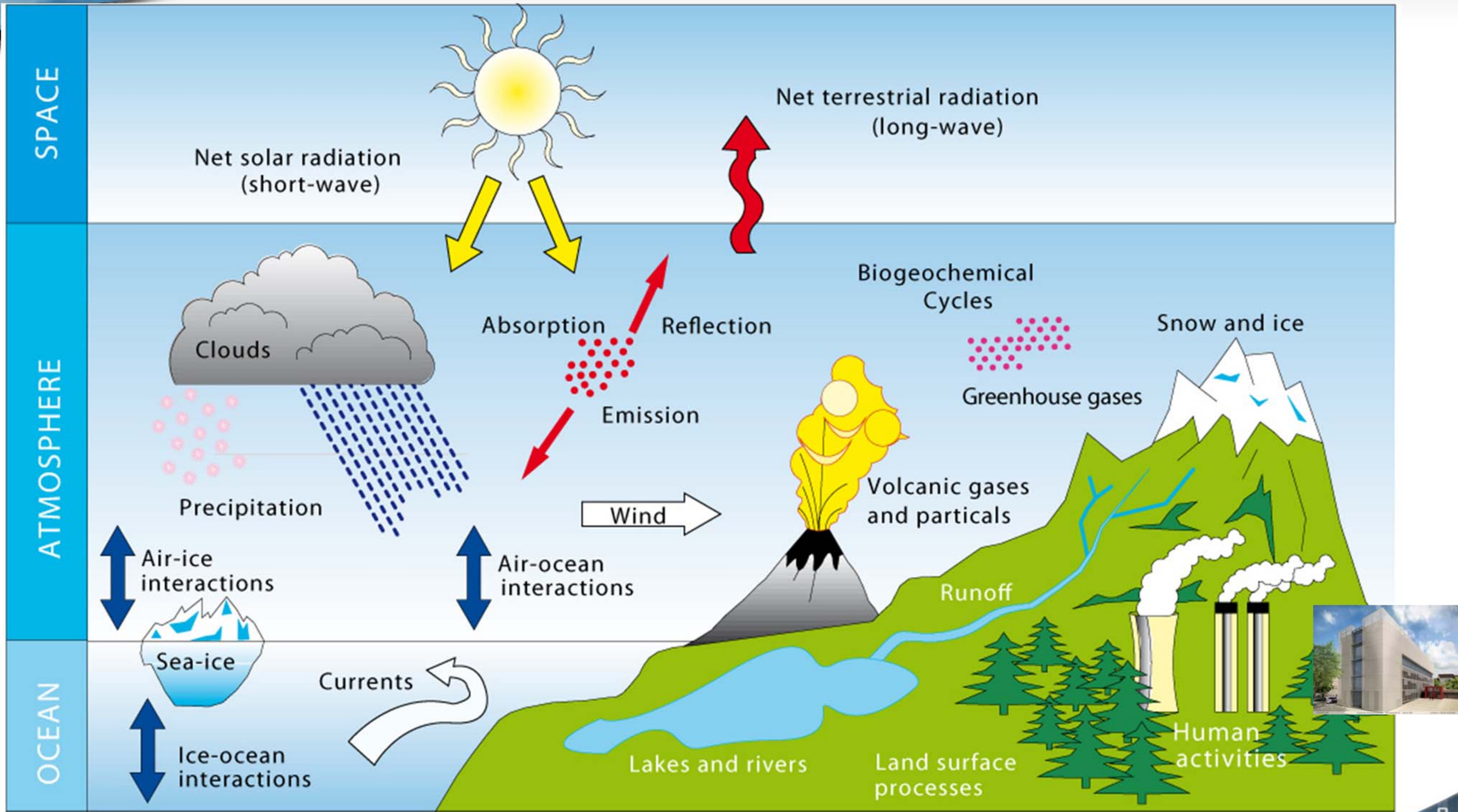
- 100 PB storage capacity
- 90 tape drives
- HPSS HSM system

Mission

DKRZ – to provide high performance computing platforms, sophisticated and high capacity data management, and superior service for premium climate science

- Operated as a non-profit company with Max-Planck-Society as principal share holder
- 60+ staff

Climate Modelling

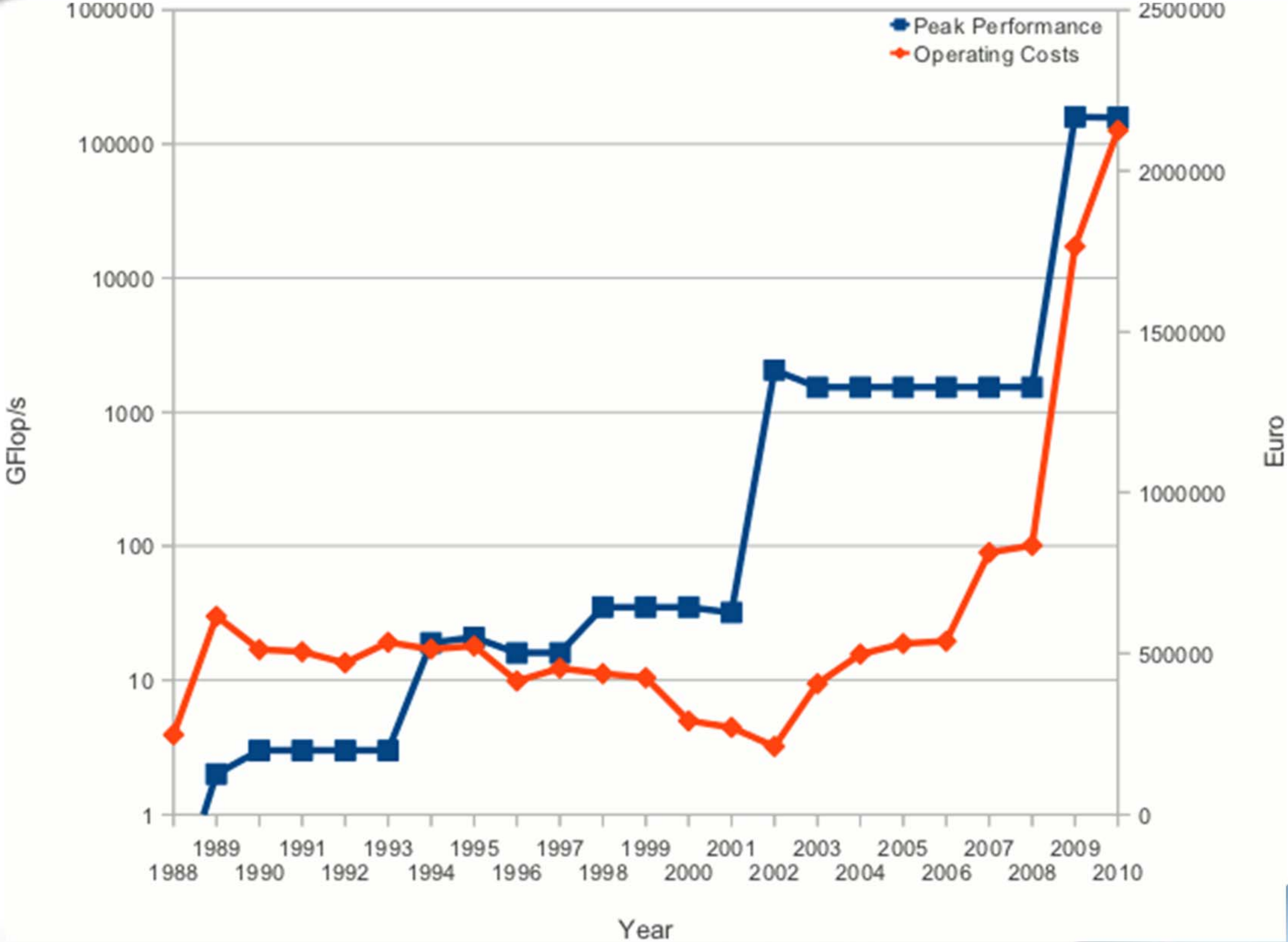


Energy Costs at DKRZ

- 2 MW for computer, storage, cooling, building
- Annual budget for power >2 M€
- Currently we use certified renewable energy
  - Otherwise ca. 10.000t CO_2/y

- High performance compute centres: 1-10 MW
- Energy costs become limiting factor for HPC usage

Energy Cost History at DKRZ



Business Model at Google

Business Model at DKRZ

Energy Costs for Science

5th IPCC status report:

- German part uses ca. 30M corehours at DKRZ
- DKRZ offers ca. 60M corehours/y
- Energy costs for the German IPCC contribution: ca. 1 M€
  - **9.000.000 kWh to solution** with DKRZ´s Blizzard system
  - 4.500.000 kg of CO_2 with regular German electricity

Climate researchers should predict the climate change…

… and not produce it!

Total Costs at DKRZ

Total costs of ownership (TCO)

- Building: 25 M€ / 25 y
- Computer and storage: 36 M€ / 5 y
- Electricity: 2 M€/y
- Others costs at DKRZ: 6 M€/y

TCO of DKRZ per year:	approximately 16 M€
Processor hours per year:	approximately 60 M
Prize per processor hour:	about 40 Cent

Total Costs for Science at DKRZ

TCO of DKRZ per year: approximately 16 M€

Publications per year: let´s assume 400

Mean price per publication: 40.000 €

- Could be justifiable for climate science
- What about astro physics and e.g. galaxy collisions ?

The Exascale Era

The Exascale Era

In approximately 2019 we will hit the next improvement of factor 1000

Same procedure as every ten years?

- Just more powerful computers? Exaflops
- Just more disks? Exabytes

From Petascale to Exascale: evolution or revolution?

Terascale to Petascale: evolution

– Just more of MPI-Fortran/C/C++

Expected Systems Architecture

Systems	2009	2018	Difference Today & 2018
System peak	2 Pflop/s	1 Eflop/s	O(1000)
Power	6 MW	~20 MW	
System memory	0.3 PB	32-64 PB [.03 Bytes/Flop]	O(100)
Node performance	125 GF	1, 2 or 15 TF	O(10)-O(100)
Node memory BW	25 GB/s	2-4 TB/s [.002 Bytes/Flop]	O(100)
Node concurrency	12	O(1k) or O(10k)	O(100)-O(1000)
Total node interconnect BW	3.5 GB/s	200-400 GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10)-O(100)
Total concurrency	225,000	O(billion) [O(10) to O(100) for latency hiding]	O(10000)
Storage	15 PB	500-1000 PB (>10x system memory is min)	O(10)-O(100)
IO	0.2 TB	60 TB/s (how long to drain the machine)	O(100)
MTTI	days	O(1 day)	- O(10)

The Exascale Revolution

Some sort of disruptiveness

- Many more processors
- More diverse hardware (e.g. GPUs)
- More levels in memory hierarchy
- Mandatory energy efficiency

TCO Considerations

- Computer in the range of 100 M€
- Power in the range of 20 M€/y (= 100 M€ in 5 years)
- I.e. 40 M€/y + staff

Exascale Science

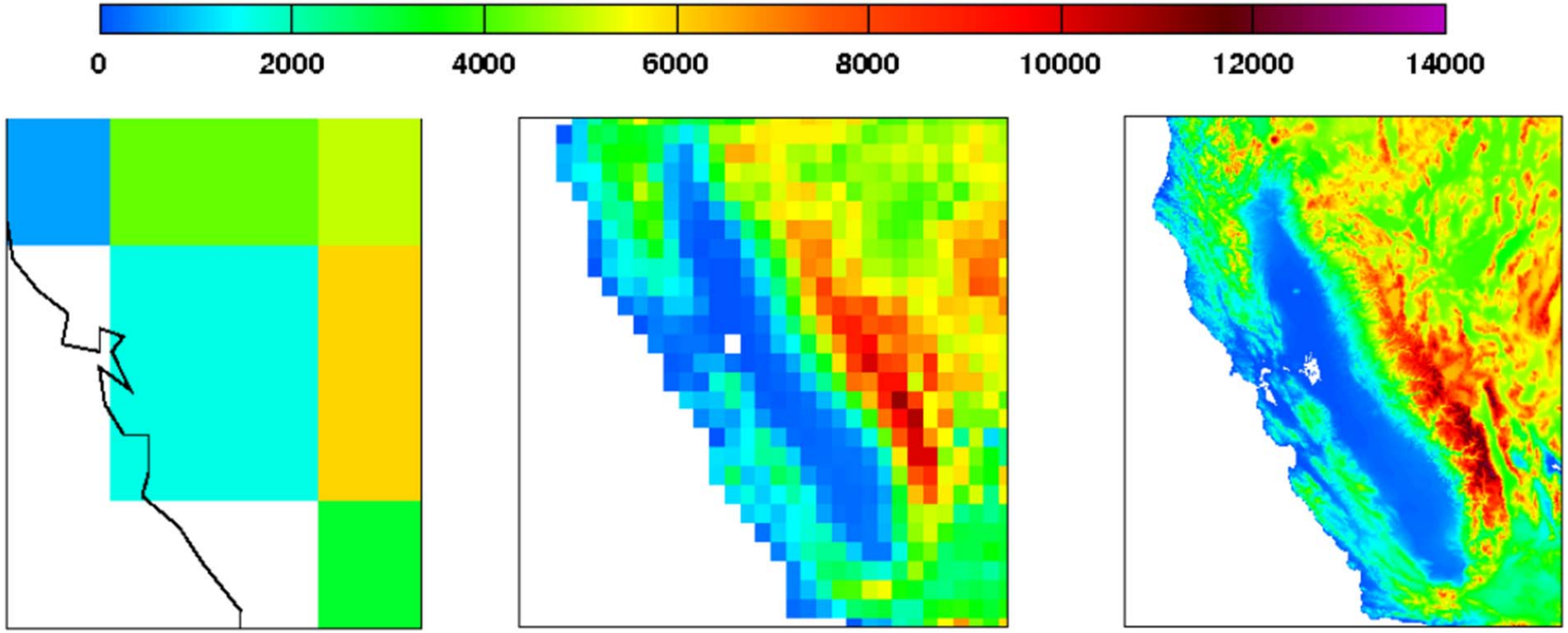
The usual "finally-we-can-do" suspects

- Biology: simulate the human brain
- Particle physics: find the Higgs Boson
- Medical science: eliminate cancer, Alzheimer etc.
- Astrophysik: understand galaxy collisions
- ...

However, what we learn here:

Modern science depends on high performance computing!

Finally: cloud computing



200km
Typical resolution of
IPCC AR4 models

25km
Upper limit of climate models
with cloud parameterizations

1km
Cloud system resolving models
are a transformational change

Power Consumption Development

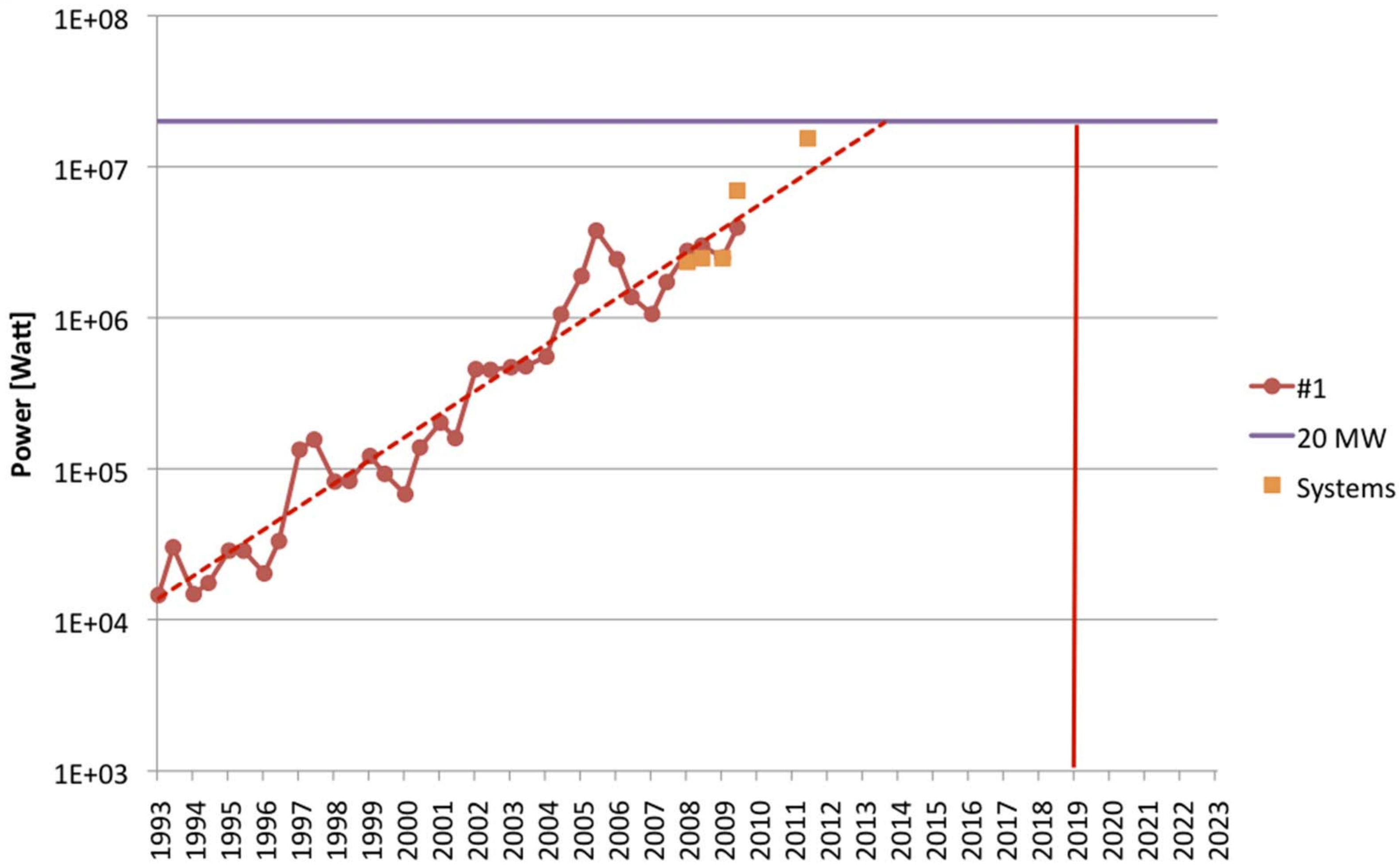


Figure by courtesy of ZIH Dresden

Power Efficiency Development

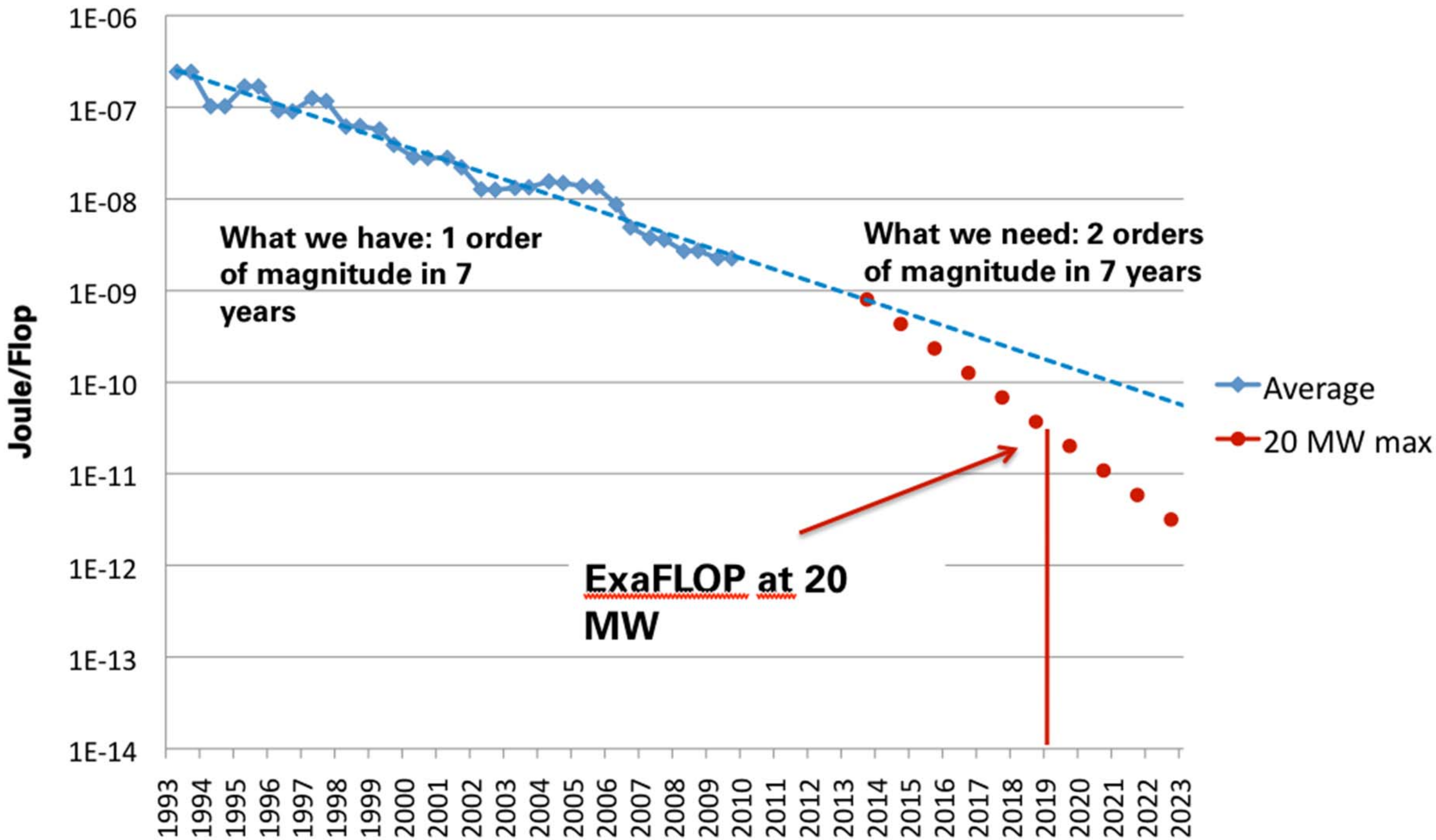


Figure by courtesy of ZIH Dresden

Costs of Science in Future

Problems for exascale HPC based science

- Power consumption might be too high and nobody will be willing to pay for it

Consequences / Requirements

- Look for higher energy efficiency in all components

What, if we are not successful?

- Will harm the Western science and engineering productivity

Research and Development

Goal: sustained HPC-based science and engineering

EESI – European Exascale Software Initiative

- WG 4.2 Software Ecosystems
  - Subtopic Power Management
    - Works on concepts for research on energy efficiency

In general much research in Europe on energy efficiency

Energie Efficiency Research

Levels of Activity

Hardware – Software – Brainware

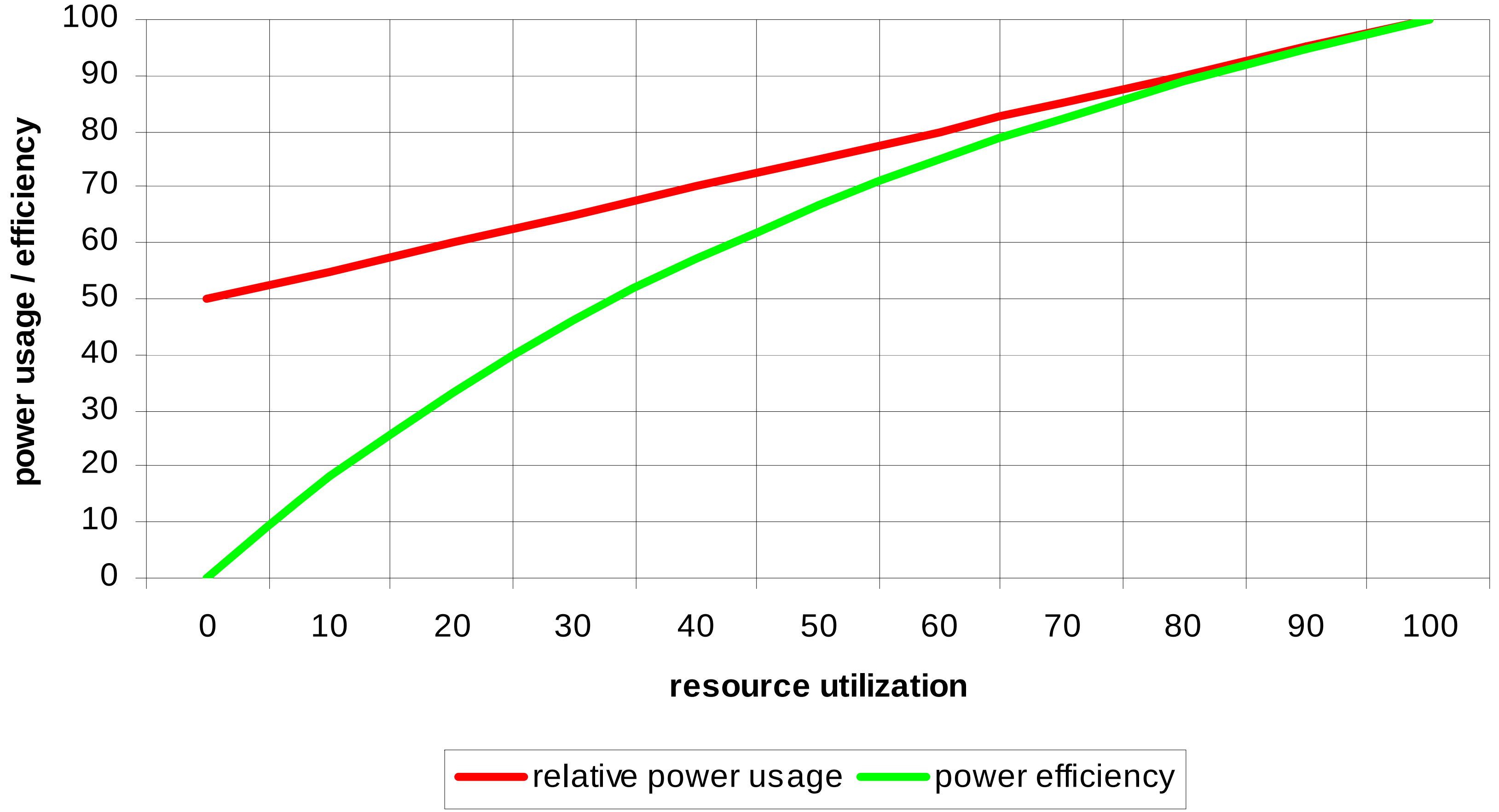
Matériel – Logiciel – Cervelle

Energy Efficient Hardware

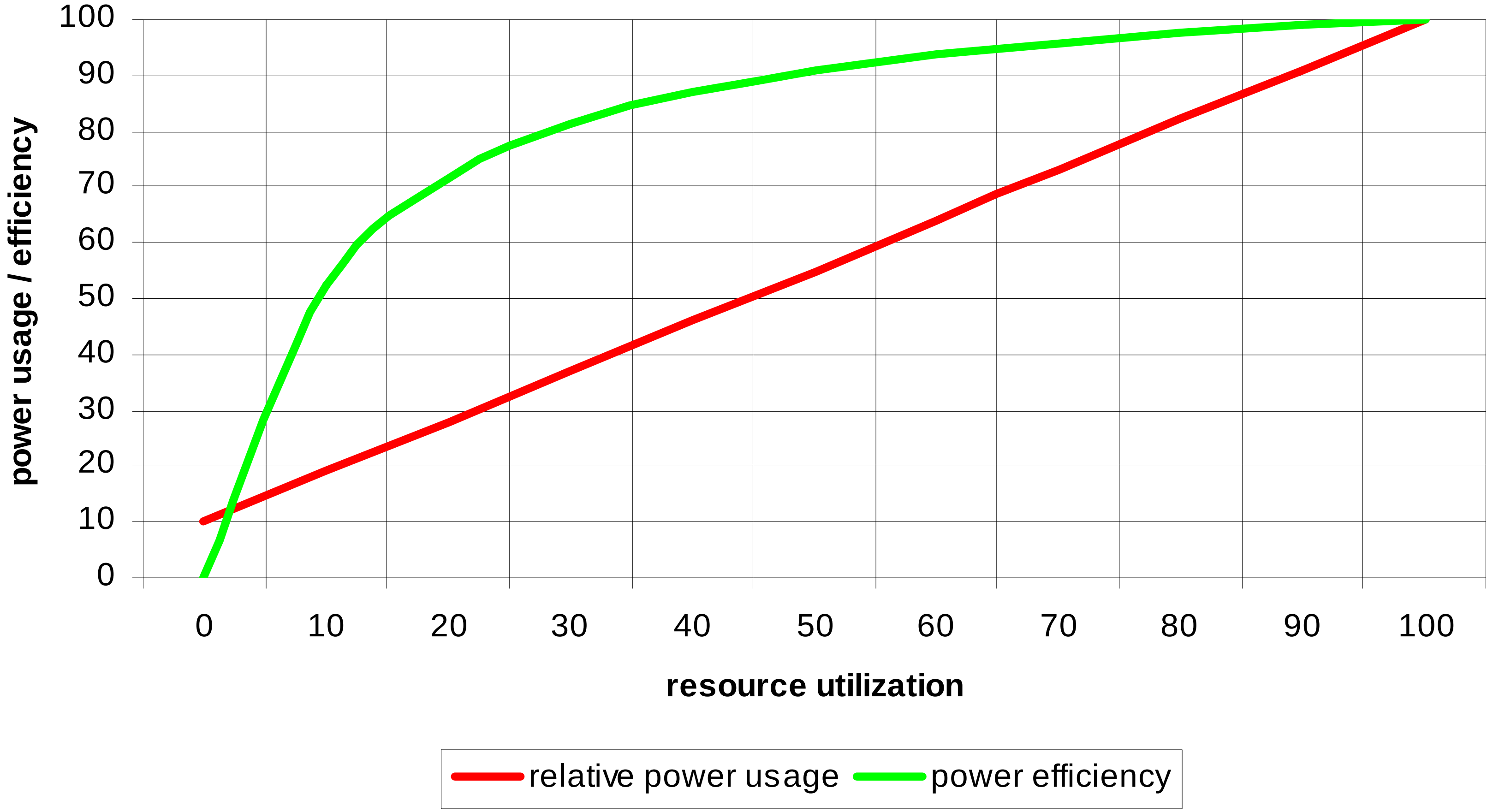
Progress at all levels is needed

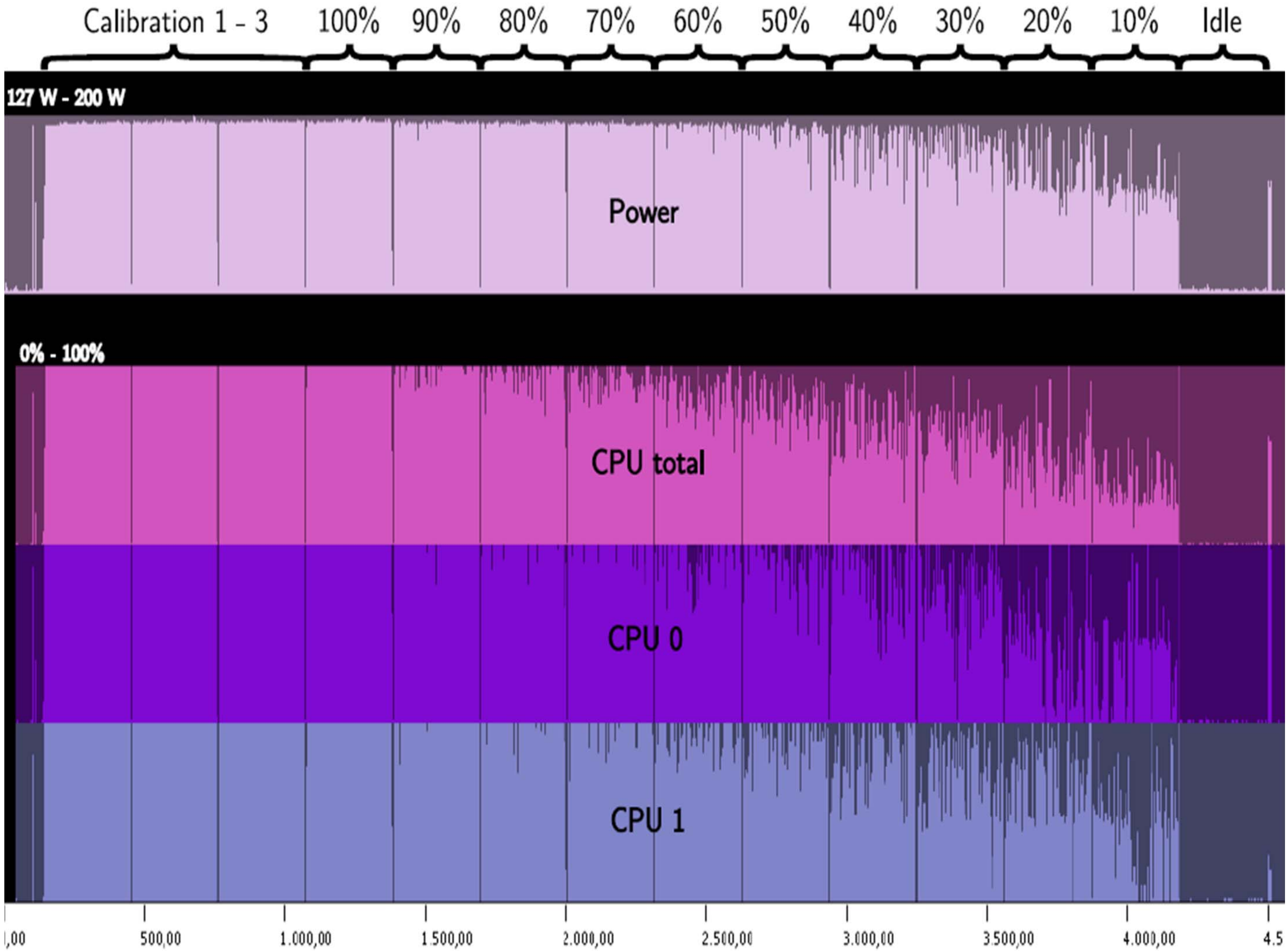
- Lower energy consumption in all parts
  - We see progress with semiconductor technology
  - We see other technologies at the horizon
    - Carbon nano tubes
    - Biocomputers
    - Quantum computers
- Power proportionality is needed
  - High consumption with high load
  - Low consumption with low load

Poor Power Proportionality



High Power Proportionality




Calibration 1 - 3
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
Idle
127 W - 200 W
Power
0% - 100%
CPU total
CPU 0
CPU 1
500,00
1.000,00
1.500,00
2.000,00
2.500,00
3.000,00
3.500,00
4.000,00
4.5

Power Proportional Hardware

High energy-proportionality

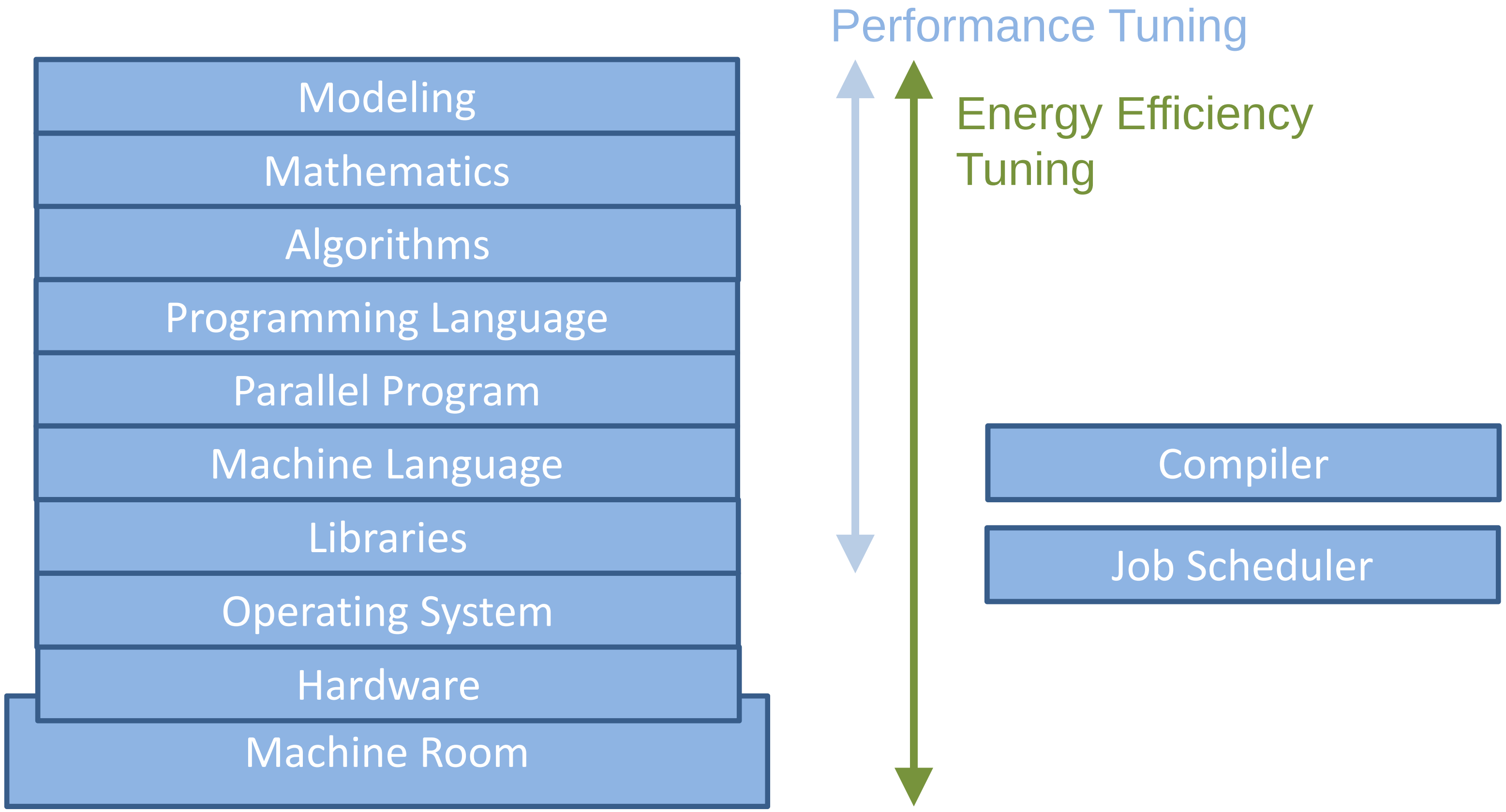
- CPUs, in particular for mobile and embedded systems
- use energy saving modes
- smooth mode switching

Poor energy-proportionality

- disk drives
- network components
- DRAM

mode switching with reactivation penalties

Abstraction Levels



Energy Efficiency Research

- Modeling
  - Which energy consumption can be seen with which hardware for which application ?
  - Which is the optimal system for an application ?

    HPC system / Grid / Cloud
  - How much energy does it take to move my application there ?
- Simulation
  - How behaves environment A compared to environment B ?
  - How behaves a rearranged software ?
- Measurement
  - Where can I measure what (hardware/software) ?

Energy Efficiency Research...

- Evaluation
  - Visualize and understand measurements
  - Automatic analysis of energy bottlenecks ?
- Improved Concepts
  - Facility management / computer hardware / operating system / middle-ware / programming / job and data scheduling
- Benchmarking

Research at University Hamburg

- Energy Efficient Cluster Computing (eeClust)
  - Analyze parallel programs
    - Trace based analysis
  - Find phases of resource inactivity
  - Switch resources into power saving modes during these phases
    - Instrumentation entered into source code

Goal: switch off all unused hardware and minimize reactivation penalty

Brainware

Let us have a commercial look at scientific applications

- They have high costs to develop them (human resources)
- They have high costs to run them (electricity)
- Some have high costs to save the results (disks, tapes)

Electricity costs are an overproportional high factor

- Use better hardware and software to reduce costs
- Use brainware to reduce the runtime and thus reduce costs

Brainware...

Example IPCC AR5 production runs

Tune program and save 10% runtime

- Saves 900.000 kWh
- Saves 100.000€
- Is 1,5 years of a skilled tuning specialist at DKRZ

Real examples are e.g. available from

– HECToR: UK National Supercomputing Service

- Success stories on code tuning and corresponding budget savings

The Future of Computational Science and Engineering

Future Architectures

A few Exascale systems for capability computing

- Difficult to use efficiently
- Expensive to operate (> 50 M€ annual budget)

Grid Infrastructures

- Based on existing concepts
- Uses tier-1 compute centres

Cloud infrastructures

- All sorts of services will be offered and used
- Commercial and non-commercial providers
- Can offer good prices for computing and storage

Future Usage Concepts

Map application to appropriate environment

What is appropriate?

- Cheap to transfer the application
- Cheap to execute the application

Transfer costs for code and data must be considered

- Model of resource usage defined by the applications
- Data intensive applications are critical

Energy aware scheduling

- Models and solutions are available

Future Policy

Objective: minimize kWh-to-solution

- For more science in a shorter time
- For a cheaper science
- For a greener science

What do we need

- Adapted funding systems

  More people, less iron
- Education of computer scientists

  Currently there are not enough

Paradigm Shift

from

"time to solution"

to

"**kWh to solution**"

for a more

economical & ecological

science

Perhaps see you again at...

EnA-HPC 2011

Second International Conference on

Energy-Aware High Performance Computing

September 7-9, 2011

Hamburg

www.ena-hpc.org